Efficient certication of numeric solutions to eigenproblems

Joris van Der Hoeven, Bernard Mourrain

Efficient certification of numeric solutions to eigenproblems

Joris van der Hoeven
Laboratoire d'informatique, UMR 7161 CNRS
Campus de l'École polytechnique
1, rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing, CS35003
91120 Palaiseau
France
Email: vdhoeven@lix.polytechnique.fr

Bernard Mourrain
Inria Sophia Antipolis Méditerranée
AROMATH
2004 route des Lucioles
06902 Sophia Antipolis
France
Email: Bernard.Mourrain@inria.fr

August 30, 2017

In this paper, we present an efficient algorithm for the certification of numeric solutions to eigenproblems. The algorithm relies on a mixture of ball arithmetic, a suitable Newton iteration, and clustering of eigenvalues that are close.

Keywords: Ball arithmetic, interval arithmetic, reliable computing, computable analysis

A.M.S. subject classification: 65G20, 03F60, 65F99, 37-04

1. Introduction

Let $\mathbb{F}$ be the set of floating point numbers for a fixed precision and a fixed exponent range. We will denote $\mathbb{F}^{\geqslant} = \{x \in \mathbb{F} : x \geqslant 0\}$. Consider an $n \times n$ matrix $M \in \mathbb{F}[\mathrm{i}]^{n \times n}$ with complex floating entries. The numeric *eigenproblem* associated to M is to compute a transformation matrix $T \in \mathbb{F}[\mathrm{i}]^{n \times n}$ and a diagonal matrix $D \in \mathbb{F}[\mathrm{i}]^{n \times n}$ such that

$$D \;\approx\; T^{-1} M T. \tag{1}$$

The entries of D are the approximate eigenvalues and the columns of T are the approximate eigenvectors of M. In addition, we might require that T is normalized. For instance, each of the columns might have unit norm. Alternatively, the norm of the i-th column may be required to be the same as the norm of the i-th row of T^{-1}, for each i. There are several well-known algorithms for solving the numeric eigenproblem [4].

Unfortunately, (1) is only an approximate equality. It is sometimes important to have rigourous bounds for the distance between the approximate eigenvalues and/or eigenvectors and the genuine ones. More precisely, we may ask for a diagonal matrix $D_r \in (\mathbb{F}^{\geqslant})^{n \times n}$ and a matrix $T_r \in (\mathbb{F}^{\geqslant})^{n \times n}$ of radii such that there exists a matrix $T' \in \mathbb{C}^{n \times n}$ for which

$$D' \;=\; (T')^{-1} M T$$

is diagonal and

$$\begin{aligned} |D'_{i,i} - D_{i,i}| &\leqslant (D_r)_{i,i} \\ |T'_{i,j} - T_{i,j}| &\leqslant (T_r)_{i,j} \end{aligned}$$

for all i, j. This task will be called the *certification problem* of the numeric solution (D, T) to the eigenproblem for M.

It will be convenient to rely on *ball arithmetic* [6, 8], which is a systematic technique for bound computations. When computing with complex numbers, ball arithmetic is more accurate than more classical interval arithmetic [11, 1, 12, 7, 9, 15], especially in multiple precision contexts. We will write $\mathbb{B} = \mathcal{B}(\mathbb{F}[\mathrm{i}], \mathbb{F}^{\geqslant})$ for the set of balls $\boldsymbol{z} = \mathcal{B}(z_c, z_r) = \{z \in \mathbb{C}: |z - z_c| \leqslant z_r\}$ with centers z_c in $\mathbb{F}[\mathrm{i}]$ and radii z_r in $\mathbb{F}^{\geqslant}$. In a similar way, we may consider matricial balls $\boldsymbol{M} = \mathcal{B}(M_c, M_r) \in \mathcal{B}(\mathbb{F}[\mathrm{i}]^{n\times n}, (\mathbb{F}^{\geqslant})^{n\times n})$: given a center matrix $M_c \in \mathbb{F}[\mathrm{i}]^{n\times n}$ and a radius matrix $M_r \in (\mathbb{F}^{\geqslant})^{n\times n}$, we have

$$\boldsymbol{M} \;=\; \mathcal{B}(M_c, M_r) \;=\; \{M \in \mathbb{C}^{n\times n}: \forall i, j, |(M_c)_{i,j} - M_{i,j}| \leqslant (M_r)_{i,j}\}.$$

Alternatively, we may regard $\mathcal{B}(M_c, M_r)$ as the set of matrices in $\mathbb{B}^{n\times n}$ with ball coefficients:

$$\mathcal{B}(M_c, M_r)_{i,j} \;=\; \mathcal{B}((M_c)_{i,j}, (M_r)_{i,j}).$$

Standard arithmetic operations on balls are carried out in a reliable way. For instance, if $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{B}$, then the computation of the product $\boldsymbol{w} = \boldsymbol{u}\,\boldsymbol{v}$ using ball arithmetic has the property that $u\,v \in \boldsymbol{w}$ for any $u \in \boldsymbol{u}$ and $v \in \boldsymbol{v}$. Given a ball $\boldsymbol{z} \in \mathbb{B}$, it will finally be convenient to write $\lfloor \boldsymbol{z} \rfloor \in \mathbb{F}^{\geqslant}$ and $\lceil \boldsymbol{z} \rceil \in \mathbb{F}^{\geqslant}$ for certified lower and upper bounds of $|\boldsymbol{z}|$ in $\mathbb{F}^{\geqslant}$.

In the language of ball arithmetic, it is natural to allow for small errors in the input and replace the numeric input $M \in \mathbb{F}[\mathrm{i}]^{n\times n}$ by a ball input $\mathcal{B}(M_c, M_r) \in \mathbb{B}^{n\times n}$. Then we may still compute a numeric solution

$$D_c \;\approx\; T_c^{-1}\, M_c\, T_c, \tag{2}$$

for the eigenproblem associated to the center M_c. Assume that the matrices in $\mathcal{B}(M_c, M_r)$ are all diagonalizable. The generalized *certification problem* now consists of the computation of a diagonal matrix $D_r \in (\mathbb{F}^{\geqslant})^{n\times n}$ and a matrix $T_r \in \mathbb{F}[\mathrm{i}]^{n\times n}$ such that, for every $M \in \mathcal{B}(M_c, M_r)$, there exist $D \in \mathcal{B}(D_c, D_r)$ and $T \in \mathcal{B}(T_c, T_r)$ with

$$D \;=\; T^{-1}\, M\, T.$$

In absence of multiple eigenvalues, known algorithms for solving this problem such as [17, 14] proceed by the individual certification of each eigenvector, which results in an $O(n^4)$ running time.

Extensions to a cluster of eigenvalues and the corresponding eigenvectors have been considered in [3, 16], with similar $O(n^4)$ complexity bounds. Fixed points theorem based on interval arithmetic are used to prove the existence of a matrix with a given Jordan block in the matrix interval domain. Such an approach has been exploited for the analysis of multiple roots in [5, 13]. A test that provides an enclosing of all the eigenvalues has been proposed in [10]. Its certification relies on interval and ball arithmetics. The complexity of the test is in $O(n^3)$ but no iteration converging to the solution of the eigenproblem is described.

In this paper, we present a new algorithm of time complexity $O(n^3)$ for certifying and enclosing clusters of eigenvectors and eigenvalues in a single step. We also provide an iterative procedure that converges geometrically to clusters of solutions. This convergence is quadratic in the case of single eigenvalues. Our algorithm extends a previous algorithm from [6] to the case of multiple eigenvalues. This yields an efficient test for *approximate eigenvalues* in the sense of the α-theory [2].

We recall that it is very unlikely that the numeric matrix $M_c \in \mathbb{F}[\mathrm{i}]^{n\times n}$ with complex floating point coefficients has multiple eigenvalues. Indeed, small perturbations of matrices with multiple eigenvalues, as induced by rounding errors, generically only have simple eigenvalues. Consequently, we may assume without loss of generality that the numeric eigenproblem (2) has a reasonably accurate solution (if necessary, we may slightly perturb M_c and increase M_r accordingly). Using ball arithmetic, it is straightforward to compute the matricial ball

$$\mathcal{B}(N_c, N_r) \;=\; \mathcal{B}(T_c, 0)^{-1}\, \mathcal{B}(M_c, M_r)\, \mathcal{B}(T_c, 0).$$

If our numerical algorithm is accurate, then the non diagonal entries of $\mathcal{B}(N_c, N_r)$ tend to be small, whence $\mathcal{B}(N_c, N_r)$ can be considered as a small perturbation of a diagonal matrix. If we can estimate how far eigenvalues and eigenvectors of diagonal matrices can drift away under small perturbations, we thus obtain a solution to the original certification problem.

Section 2 introduces notations. In Section 3, we perform a detailed study of the eigenproblem for small perturbations M of diagonal matrices. We exhibit a Newton iteration for finding the solutions. This iteration has quadratic convergence in the absence of multiple eigenvalues and is also an efficient tool for doubling the precision of a solution. However, in the case of multiple eigenvalues, the eigenproblem is ill-posed. Indeed, by a well-known observation, *any* vector occurs as the eigenvector of a small perturbation of the 2×2 identity matrix. The best we can hope for is to group eigenvectors with close eigenvalues together in "clusters" (see also [16]) and only require $T^{-1}\, M\, T$ to be block diagonal. For this reason, we present our Newton iteration in a sufficiently general setting which encompasses block matrices. We will show that the iteration still admits geometric convergence for sufficiently small perturbations and that the blockwise certification is still sufficient for the computation of rigourous error bounds for the eigenvalues. In Section 4, we will present explicit algorithms for clustering and the overall certification problem.

2. Notations

2.1. Matrix norms

Throughout this paper, we will use the max norm for vectors and the corresponding matrix norm. More precisely, given a vector $v \in \mathbb{C}^n$ and an $n \times n$ matrix $M \in \mathbb{C}^{n\times n}$, we set

$$\begin{aligned} \|v\| &= \max\{|v_1|, ..., |v_n|\} \\ \|M\| &= \max_{\|v\|=1} \|Mv\|. \end{aligned}$$

For a second matrix $N \in \mathbb{C}^{n\times n}$, we clearly have

$$\begin{aligned} \|M+N\| &\leqslant \|M\| + \|N\| \\ \|MN\| &\leqslant \|M\|\, \|N\|. \end{aligned}$$

Explicit machine computation of the matrix norm is easy using the formula

$$\|M\| \;=\; \max\{|M_{i,1}| + \cdots + |M_{i,n}| : 1 \leqslant i \leqslant n\}. \tag{3}$$

In particular, when changing certain entries of a matrix M to zero, its matrix norm $\|M\|$ can only decrease.

2.2. Clustering

Assume that we are given a partition

$$\{1, ..., n\} \;=\; I_1 \amalg \cdots \amalg I_p. \tag{4}$$

Such a partition will also be called a *clustering* and denoted by I. Two indices i, j are said to belong to the same *cluster* if there exists a k with $\{i, j\} \subseteq I_k$ and we will write $i \sim j$. Two entries $M_{i,j}$ and $M_{i',j'}$ of a matrix $M \in \mathbb{C}^{n\times n}$ are said to belong to the same *block* if $i \sim j$ and $i' \sim j'$. We thus regard M as a generalized block matrix, for which the rows and columns of the blocks are not necessarily contiguous inside M.

A matrix $M \in \mathbb{C}^{n\times n}$ is said to be *block diagonal* (relative to the clustering) if $M_{i,j} = 0$ whenever $i \nsim j$. Similarly, we say that M is *off block diagonal* if $M_{i,j} = 0$ whenever $i \sim j$. For a general $M \in \mathbb{C}^{n\times n}$, we define its block diagonal and off block diagonal projections $\Delta(M) = \Delta^I(M)$ and $\Omega(M) = \Omega^I(M)$ by

$$\Delta(M)_{i,j} = \begin{cases} M_{i,j} & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases} \qquad \Omega(M)_{i,j} = \begin{cases} 0 & \text{if } i \sim j \\ M_{i,j} & \text{otherwise} \end{cases}$$

By our observation at the end of section 2.1, we have

$$\begin{aligned} \|\Delta(M)\| &\leqslant \|M\| \\ \|\Omega(M)\| &\leqslant \|M\|. \end{aligned}$$

For the *trivial clustering* $I_k = \{k\}$, the matrices $\Delta(M)$ and $\Omega(M)$ are simply the diagonal and off diagonal projections of M. In that case we will also write $\Delta^* = \Delta$ and $\Omega^* = \Omega$.

2.3. Diagonal matrices

Below, we will study eigenproblems for perturbations of a given diagonal matrix

$$D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}. \tag{5}$$

It follows from (3) that the matrix norm $\mu = \|D\|$ of a diagonal matrix D is given by

$$\mu = \max\{|\lambda_1|, ..., |\lambda_n|\}.$$

It will also be useful to define the *separation number* $\sigma^* = \sigma^*(D)$ by

$$\sigma^* = \min\{|\lambda_i - \lambda_j| : i \neq j\}.$$

More generally, given a clustering as in the previous subsection, we also define the *block separation number* $\sigma = \sigma(D) = \sigma^I(D)$ by

$$\sigma = \min\{|\lambda_i - \lambda_j| : i \nsim j\}$$

This number σ remains high if the clustering is chosen in such a way that the indices i, j of any two "close" eigenvalues λ_i and λ_j belong to the same cluster. In particular, if $\sigma > 0$, then $\lambda_i = \lambda_j$ implies $i \sim j$.

3. EIGENPROBLEMS FOR PERTURBED DIAGONAL MATRICES

3.1. The linearized equation

Let D be a diagonal matrix (5). Given a small perturbation

$$M = D + H$$

of D, where H is an off diagonal matrix, the aim of this section is to find a small matrix $E \in \mathbb{C}^{n\times n}$ for which

$$M' \;=\; (1+E)^{-1}\, M\, (1+E)$$

is block diagonal. In other words, we need to solve the equation

$$\Omega((1+E)^{-1}\,(D+H)\,(1+E)) \;=\; 0.$$

When linearizing this equation in E and H, we obtain

$$\Omega([D,E]+H) \;=\; 0.$$

If E is strongly off diagonal, then so is $[D,E]$, and the equation further reduces to

$$[D,E] \;=\; -\Omega(H).$$

This equation can be solved using the following lemma:

Lemma 1. *Given a matrix $A \in \mathbb{C}^{n\times n}$ and a diagonal matrix D with entries $\lambda_1, ..., \lambda_n$, let $B = \Phi(D,A) \in \mathbb{C}^{n\times n}$ be the strongly off diagonal matrix with*

$$B_{i,j} \;=\; \begin{cases} 0 & \text{if } i \sim j \\ \frac{A_{i,j}}{\lambda_j - \lambda_i} & \text{otherwise} \end{cases}$$

Then $\|B\| \leqslant \sigma^{-1}\,\|A\|$ and

$$[D,B] \;=\; -\Omega(A). \tag{6}$$

Proof. The inequality follows from (3) and the definition of σ. One may check (6) using a straightforward computation. □

3.2. The fundamental iteration

In view of the lemma, we now consider the iteration

$$(D,H) \;\longmapsto\; (D',H'),$$

where

$$\begin{aligned} E &= \Phi(D,H) \\ M' &= (1+E)^{-1}(D+H)\,(1+E) \\ D' &= \Delta^*(M') \\ H' &= \Omega^*(M') \end{aligned}$$

In order to study the convergence of this iteration, we introduce the quantities

$$\begin{aligned} \mu &= \|D\| & \mu' &= \|D'\| \\ \sigma &= \sigma(D) & \sigma' &= \sigma(D') \\ \eta_1 &= \|\Delta(H)\| & \eta_1' &= \|\Delta(H')\| \\ \eta_2 &= \|\Omega(H)\| & \eta_2' &= \|\Omega(H')\| \end{aligned}$$

$$\alpha \;=\; \min\left\{\frac{\sigma}{6\,\mu}, \frac{1}{4}\right\}.$$

LEMMA 2. *For $\delta \in (0,1]$, assume that*

$$\begin{aligned} \eta_1 + \eta_2 &\leqslant \alpha\,\delta\,\mu \\ \eta_2 &\leqslant \alpha\,\delta\,\sigma. \end{aligned}$$

Then $\|D' - D\| \leqslant \delta\,\eta_2$ and

$$\begin{aligned} \mu' &\leqslant \mu + \delta\,\eta_2 \\ \sigma' &\geqslant \sigma - 2\,\delta\,\eta_2 \\ \eta_1' &\leqslant \eta_1 + \delta\,\eta_2 \\ \eta_2' &\leqslant \delta\,\eta_2. \end{aligned}$$

Proof. We have

$$\begin{aligned} M' - D &= H + [D, E] + R \\ &= \Delta(H) + R, \end{aligned}$$

where

$$R = E^2\,(1+E)^{-1}\,(D+H)\,(1+E) - E\,(D+H)\,E + [H, E].$$

Setting $\varepsilon = \|E\| \leqslant \sigma^{-1}\,\eta_2 \leqslant \alpha\,\delta \leqslant \frac{1}{4}$, the remainder R is bounded by

$$\begin{aligned} \|R\| &\leqslant \varepsilon^2 \frac{1}{1-\varepsilon}\,(1+\alpha\,\delta)\,\mu\,(1+\varepsilon) + \varepsilon\,(1+\alpha\,\delta)\,\mu\,\varepsilon + 2\,(\eta_1+\eta_2)\,\varepsilon \\ &= \frac{2\,\varepsilon^2}{1-\varepsilon}\,(1+\alpha\,\delta)\,\mu + 2\,(\eta_1+\eta_2)\,\varepsilon \\ &\leqslant (4\,\varepsilon\,\mu + 2\,\alpha\,\delta\,\mu)\,\varepsilon \\ &\leqslant 6\,\alpha\,\delta\,\mu\,\sigma^{-1}\,\eta_2 \\ &\leqslant \delta\,\eta_2. \end{aligned}$$

Consequently,

$$\begin{aligned} \|D' - D\| &= \|\Delta^*(M' - D)\| = \|\Delta^*(R)\| \\ &\leqslant \|R\| \leqslant \delta\,\eta_2 \\ \eta_1' &= \|\Delta(H')\| = \|\Omega^*(\Delta(M'))\| = \|\Omega^*(\Delta(H+R))\| \\ &\leqslant \|H+R\| \leqslant \eta_1 + \delta\,\eta_2 \\ \eta_2' &= \|\Omega(H')\| = \|\Omega(M')\| = \|\Omega(R)\| \\ &\leqslant \delta\,\eta_2. \end{aligned}$$

The inequalities $\mu' \leqslant \mu + \delta\,\eta_2$ and $\sigma' \geqslant \sigma - 2\,\delta\,\eta_2$ follow from $\|D' - D\| \leqslant \delta\,\eta_2$. □

3.3. Convergence of the fundamental iteration

THEOREM 3. *Assume that*

$$\begin{aligned} \eta_1 + \eta_2 &\leqslant \tfrac{1}{8}\,\alpha\,\mu \\ \eta_2 &\leqslant \tfrac{1}{8}\,\alpha\,\sigma. \end{aligned}$$

Then the sequence

$$(D, H), (D', H'), (D'', H''), \ldots$$

converges geometrically to a limit $(M^{(\infty)}, H^{(\infty)})$ *with* $\|D^{(\infty)} - M\| \leqslant \eta_2$ *and* $\|H^{(\infty)}\| \leqslant \eta_1 + \eta_2$. *The matrix* $D^{(\infty)} + H^{(\infty)}$ *is block diagonal and there exists a matrix* $\hat{E}$ *with* $\|\hat{E}\| \leqslant 3\,\sigma^{-1}\,\eta_2$, *such that*

$$D^{(\infty)} + H^{(\infty)} \;=\; (1+\hat{E})^{-1}\,(D+H)\,(1+\hat{E}).$$

Proof. Let $(D^{(i)}, H^{(i)})$ stand for the i-th fundamental iterate of (D, H) and $E^{(i)} = \Phi(H^{(i)}, D^{(i)})$. Denote $\mu^{(i)} = \|D^{(i)}\|$, $\sigma^{(i)} = \sigma(D^{(i)})$, $\eta_1^{(i)} = \|\Delta(H^{(i)})\|$ and $\eta_2^{(i)} = \|\Omega(H^{(i)})\|$. Let us show by induction over i that

$$\begin{aligned}
\|D^{(i)} - D\| &\leqslant (1 - \tfrac{1}{2^i})\,\eta_2 \\
\mu^{(i)} &\leqslant \mu + (1 - \tfrac{1}{2^i})\,\eta_2 \\
\sigma^{(i)} &\geqslant \tfrac{1}{2}\,(1 + \tfrac{1}{2^i})\,\sigma \\
\eta_1^{(i)} &\leqslant \eta_1 + (1 - \tfrac{1}{2^i})\,\eta_2 \\
\eta_2^{(i)} &\leqslant \tfrac{1}{2^i}\,\eta_2.
\end{aligned}$$

This is clear for $i = 0$. Assume that the induction hypothesis holds for a given i and let

$$\alpha^{(i)} \;=\; \min\left\{\frac{\sigma^{(i)}}{6\,\mu^{(i)}}, \frac{1}{4}\right\}$$

Since $(1 - \frac{1}{2^i})\,\eta_2 \leqslant \frac{1}{32}\,\mu$, the induction hypothesis implies

$$\begin{aligned}
\mu^{(i)} &\leqslant 2\,\mu \\
\sigma^{(i)} &\geqslant \tfrac{1}{2}\,\sigma \\
\alpha^{(i)} &\geqslant \tfrac{1}{4}\,\alpha.
\end{aligned}$$

Applying Lemma 2 for $(D^{(i)}, H^{(i)})$ and $\delta = \frac{1}{2}$, we thus find

$$\begin{aligned}
\|D^{(i+1)} - D\| &\leqslant \|D^{(i)} - D\| + \|D^{(i+1)} - D^{(i)}\| \\
&\leqslant (1 - \tfrac{1}{2^i})\,\eta_2 + \tfrac{1}{2^{i+1}}\,\eta_2 \leqslant (1 - \tfrac{1}{2^{i+1}})\,\eta_2 \\
\mu^{(i+1)} &\leqslant \mu^{(i)} + \tfrac{1}{2}\,\eta_2^{(i)} \leqslant \mu + (1 - \tfrac{1}{2^{i+1}})\,\eta_2 \\
\sigma^{(i+1)} &\geqslant \sigma^{(i)} - \tfrac{1}{2}\,\eta_2^{(i)} \geqslant \tfrac{1}{2}\,(1 + \tfrac{1}{2^i} - \tfrac{1}{2^{i+1}})\,\sigma \geqslant \tfrac{1}{2}\,(1 + \tfrac{1}{2^{i+1}})\,\sigma \\
\eta_1^{(i+1)} &\leqslant \eta_1^{(i)} + \tfrac{1}{2}\,\eta_2^{(i)} \leqslant \eta_1 + (1 - \tfrac{1}{2^{i+1}})\,\eta_2 \\
\eta_2^{(i+1)} &\leqslant \tfrac{1}{2}\,\eta_2^{(i)} \leqslant \tfrac{1}{2^{i+1}}\,\eta_2.
\end{aligned}$$

This completes the induction.

Applying the induction to the sequence starting at $D^{(i)}$, we have for every $j \geqslant 0$,

$$\|D^{(i+j)} - D^{(i)}\| \;\leqslant\; (1 - \tfrac{1}{2^{j+1}})\,\eta_2^{(i)} \leqslant (1 - \tfrac{1}{2^{j+1}})\tfrac{1}{2^{i+1}}\,\eta_2.$$

This shows that $D^{(i)}$ is a Cauchy sequence that tends to a limit $D^{(\infty)}$ with $\|D^{(\infty)} - D\| \leqslant \eta_2$. From this inequality, we also deduce that $\|D^{(\infty)} - D^{(i)}\| \leqslant \frac{1}{2^{i+1}}\ \eta_2$, so $D^{(i)}$ converges geometrically to $D^{(\infty)}$.

Moreover, for each i, we have $\varepsilon^{(i)} = \|E^{(i)}\| \leqslant \sigma^{-1}\,\eta_2^{(i)} \leqslant \frac{1}{2^i}\,\sigma^{-1}\,\eta_2$. Hence, the matrix

$$\hat{E} = (1+E^{(0)})\,(1+E^{(1)})\,(1+E^{(2)})\cdots - 1$$

is well defined, and

$$\begin{aligned}\log(1+\|\hat{E}\|) &\leqslant \log(1+\varepsilon^{(0)}) + \log(1+\varepsilon^{(1)}) + \log(1+\varepsilon^{(2)}) + \cdots \\ &\leqslant 2\,\sigma^{-1}\,\eta_2.\end{aligned}$$

We deduce that

$$\|\hat{E}\| \leqslant \mathrm{e}^{2\sigma^{-1}\eta_2} - 1 \leqslant 3\,\sigma^{-1}\,\eta_2,$$

since $\sigma^{-1}\,\eta_2 \leqslant \frac{1}{32}$.

We claim that $M^{(i)} = D^{(i)} + H^{(i)}$ converges geometrically to

$$M^{(\infty)} = (1+\hat{E})^{-1}\,M^{(0)}\,(1+\hat{E}).$$

For any matrix $M, E \in \mathbb{C}^{n\times n}$ with $\|E\| < \varepsilon < 1$, we have

$$\begin{aligned}\|(1+E)^{-1}\,M\,(1+E) - M\| &= \|ME - E\,(1+E)^{-1}\,M\,(1+E)\| \\ &\leqslant \|M\|\,(\varepsilon + \varepsilon\,(1+\varepsilon)\,\|(1+E)^{-1}\|) \\ &\leqslant \varepsilon\,\|M\|\,(1 + (1+\varepsilon)\,(1-\varepsilon)^{-1}) \\ &= \frac{2\,\varepsilon}{1-\varepsilon}\,\|M\|. \qquad (7)\end{aligned}$$

Let $\hat{E}^{(i)} = (1+E^{(i)})\,(1+E^{(i+1)})\,(1+E^{(i+2)})\cdots - 1$. By the same arguments as above, we have $\hat{\varepsilon}_i := \|E^{(i)}\| \leqslant 3\,\sigma^{-1}\,\eta_2^{(i)} = \frac{3}{2^{i+1}}\,\sigma^{-1}\,\eta_2$. Since $M^{(\infty)} = (1+\hat{E}^{(i)})^{-1}\,M^{(i)}\,(1+\hat{E}^{(i)})$, the inequality (7) implies

$$\begin{aligned}\|M^{(\infty)} - M^{(i)}\| &\leqslant \frac{2\,\hat{\varepsilon}_i}{1-\hat{\varepsilon}_i}\,(\|D^{(i)}\| + \|H^{(i)}\|) \\ &\leqslant \frac{2\,\hat{\varepsilon}_i}{1-\hat{\varepsilon}_i}\big(\mu_i + \eta_1^{(i)} + \eta_2^{(i)}\big) \\ &\leqslant \frac{3}{2^i}\,\frac{\sigma^{-1}\,\eta_2}{1-\hat{\varepsilon}_i}\,(\mu + \eta_1 + \eta_2).\end{aligned}$$

This shows that $M^{(i)}$ converges geometrically to $M^{(\infty)}$. We deduce that the sequence $H^{(i)} = M^{(i)} - D^{(i)}$ also converges geometrically to a limit $H^{(\infty)}$ with $\|H^{(\infty)}\| \leqslant \eta_1 + \eta_2$. Since $\lim_{i\to\infty} \eta_2^{(i)} = 0$, we finally observe that $M^{(\infty)} = D^{(\infty)} + H^{(\infty)}$ is block diagonal. □

THEOREM 4. *Assume $I_k = \{k\}$ for all k. Then, under the assumptions of Theorem 3, the sequence $(D,H), (D',H'), (D'',H''), \ldots$ converges quadratically to $(D^{(\infty)}, 0)$.*

Proof. The extra assumption implies that $\eta_1^{(i)} = 0$ for all i. Let us show by induction over i that we now have

$$\eta_2^{(i)} \leqslant \frac{1}{2^{2^i-1}}\,\eta_2.$$

This is clear for $i = 0$. Assume that the result holds for a given i. Then we may apply Lemma 2 to $(D^{(i)}, H^{(i)})$ for $\delta = 2^{-2^i+1}$, and obtain

$$\begin{aligned} \eta_2^{(i+1)} &\leqslant \frac{1}{2^{2^i-1}}\, \eta_2^{(i)} \\ &\leqslant \frac{1}{2^{2^{i+1}-1}}. \end{aligned}$$

Since $\|D^{(i+1)} - D^{(i)}\| \leqslant \eta_2^{(i)}$, this establishes the quadratic convergence. □

4. Algorithms

4.1. Clustering

Let $M = D + H$ be the perturbation of a diagonal matrix (5) as in the previous section. In order to apply theorem 3, we first have to find a suitable clustering (4). For a given threshold separation δ, we will simply take the finest clustering (i.e. for which p is maximal) with the property that $|\lambda_i - \lambda_j| \leqslant \delta \Rightarrow i \sim j$. This clustering can be computed using the algorithm Cluster below.

Algorithm Cluster
Input: eigenvalues $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n \in \mathbb{B}$ and $\delta \in \mathbb{F}^{\geqslant}$
Output: the finest clustering (4) with $\lfloor \boldsymbol{\lambda}_i - \boldsymbol{\lambda}_j \rfloor \leqslant \delta \Rightarrow i \sim j$

- Let G be the graph with vertices 1, ..., n and such that i and j are connected if and only if $\lfloor \boldsymbol{\lambda}_i - \boldsymbol{\lambda}_j \rfloor \leqslant \delta$.
- Let H be the transitive closure of G.
- Let $H_1, ..., H_p$ the connected components of H.
- Let I_k be the set of vertices of H_k for each k.

4.2. Certification in the case of perturbed diagonal matrices

In order to apply theorem 3, it now remains to find a suitable threshold δ for which the conditions of the theorem hold. Starting with $\delta = 0$, we will simply increase δ to $\sigma(D)$ whenever the conditions are not satisfied. This will force the number p of clusters to decrease by at least one at every iteration, whence the algorithm terminates. Notice that the workload of one iteration is $O(n^2)$, so the total running time remains bounded by $O(n^3)$.

Algorithm DiagonalCertify
Input: a diagonal ball matrix $\boldsymbol{D} \in \mathbb{B}^{n\times n}$ with entries $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ and an off diagonal ball matrix $\boldsymbol{H} \in \mathbb{B}^{n\times n}$
Output: a clustering I and $\hat{\varepsilon} \in \mathbb{F}$ such that, for any $M \in \boldsymbol{D}$ and $H \in \boldsymbol{H}$, the conditions of theorem 3 hold and $\|\hat{E}\| \leqslant \hat{\varepsilon}$

$\delta := 0$
Repeat
  Compute the clustering I for $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n$ and δ using Cluster
  Let $\boldsymbol{\mu} := \|\boldsymbol{D}\|$, $\boldsymbol{\sigma} := \sigma^I(\boldsymbol{D})$, $\boldsymbol{\eta}_1 := \|\Delta^I(\boldsymbol{H})\|$ and $\boldsymbol{\eta}_2 := \|\Omega^I(\boldsymbol{H})\|$
  Let $\boldsymbol{\alpha} := \min\left\{\frac{\boldsymbol{\sigma}}{6\,\boldsymbol{\mu}}, \frac{1}{4}\right\}$
  If $\lceil \boldsymbol{\eta}_1 + \boldsymbol{\eta}_2 \rceil \leqslant \lfloor \frac{\boldsymbol{\alpha}\,\boldsymbol{\mu}}{8} \rfloor$ and $\lceil \boldsymbol{\eta}_2 \rceil \leqslant \lfloor \frac{\boldsymbol{\alpha}\,\boldsymbol{\sigma}}{8} \rfloor$, then return $(I, \lceil \frac{3\,\boldsymbol{\eta}_2}{\boldsymbol{\sigma}} \rceil)$
  Set $\delta := \lceil \boldsymbol{\sigma} \rceil$

4.3. Certification of approximate eigenvectors and eigenvalues

Let us now return to the original problem of certifying a numerical solution to an eigenproblem. We will denote by $\mathbb{1}_n$ the $n \times n$ matrix of which all entries are one.

Algorithm EigenvectorCertify
Input: $\boldsymbol{M} = \mathcal{B}(M_c, M_r) \in \mathbb{B}^{n\times n}$ and $T_c \in \mathbb{F}[\mathrm{i}]^{n\times n}$
such that $T_c^{-1} M_c T_c$ is approximately diagonal
Output: a clustering I and $\boldsymbol{T} = \mathcal{B}(T_c, T_r) \in \mathbb{B}^{n\times n}$ such that for any $M \in \boldsymbol{M}$,
there exists a $T \in \boldsymbol{T}$ for which $T^{-1} M T$ is block diagonal

---

Compute $\boldsymbol{D} := \mathcal{B}(T_c, 0)^{-1}\, \boldsymbol{M}\, \mathcal{B}(T_c, T_r)$
Let $(I, \varepsilon) :=$ DiagonalCertify$(\Delta^*(\boldsymbol{D}), \Omega^*(\boldsymbol{D}))$
Let $\boldsymbol{E} := \mathcal{B}(T_c, 0)\, \mathcal{B}(0, \varepsilon)\, \mathbb{1}_n$
Let $(T_r)_{i,j} := \lceil \boldsymbol{E}_{i,j} \rceil$ for all i, j
Return $(I, \mathcal{B}(T_c, T_r))$

Obviously, any eigenvalue $\lambda \in \mathbb{C}$ of a matrix $M \in \mathbb{C}^{n\times n}$ satisfies $|\lambda| \leqslant \|M\|$. We may thus use the following modification of EigenvectorCertify in order to compute enclosures for the eigenvalues of $\boldsymbol{M}$.

Algorithm EigenvalueCertify
Input: $\boldsymbol{M} = \mathcal{B}(M_c, M_r) \in \mathbb{B}^{n\times n}$ and $T_c \in \mathbb{F}[\mathrm{i}]^{n\times n}$
such that $T_c^{-1} M_c T_c$ is approximately diagonal
Output: ball enclosures $\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n \in \mathbb{B}$ for the eigenvalues of $\boldsymbol{M}$,
with the appropriate multiplicities in cases of overlapping

Compute $\boldsymbol{D} := \mathcal{B}(T_c, 0)^{-1}\, \boldsymbol{M}\, \mathcal{B}(T_c, T_r)$
Let $(I, \varepsilon) :=$ DiagonalCertify$(\Delta^*(\boldsymbol{D}), \Omega^*(\boldsymbol{D}))$
Let $\boldsymbol{\eta}_1 := \|\Delta^I(\Omega^*(\boldsymbol{D}))\|$ and $\boldsymbol{\eta}_2 := \|\Omega^I(\Omega^*(\boldsymbol{D}))\|$
For each $k \in \{1, ..., p\}$ do
  If $I_k = \{i\}$ for some i, then let $\boldsymbol{\lambda}_i := \mathcal{B}((D_c)_{i,i}, \lceil \boldsymbol{\eta}_2 \rceil)$
  Otherwise
    Let $\boldsymbol{c}$ be the barycenter of the $\boldsymbol{D}_{i,i}$ with $i \in I_k$
    Let $\boldsymbol{r}$ be the maximum of $|\boldsymbol{D}_{i,i} - c|$ for $i \in I_k$
    Let $\boldsymbol{\lambda}_i := \boldsymbol{c} + \mathcal{B}(0, \lceil \boldsymbol{r} + \boldsymbol{\eta}_1 + 2\,\boldsymbol{\eta}_2 \rceil)$ for all $i \in I_k$
Return $(\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_n)$

---

5. Possible extensions

Let $M \in \mathbb{C}^{n\times n}$ be a matrix with a (numerically) multiple eigenvalue λ. We have already stressed that it is generally impossible to provide non trivial certifications for the corresponding eigenvectors. Nevertheless, two observations should be made:

- If the eigenspace E_λ corresponding to λ has dimension 1, then small perturbations of the matrix M only induce small perturbations of λ and E_λ.
- Let F_λ denote the full invariant subspace associated to the eigenvalue λ (or all eigenvalues in the cluster of λ). Then small perturbations of M only induce small perturbations of λ and F_λ.

More precisely, in these two cases, we may search for ball enclosures for orthonormal bases of the vector spaces E_λ resp. F_λ, which do not contain the zero vector.

When considering the numeric solution (1) of the eigenproblem for M, the column vectors which generate F_λ are usually far from being orthogonal. Orthonormalization can only be done at the expense of making $T^{-1}\,M\,T$ only upper triangular. Moreover, the orthogonalization implies a big loss of accuracy, which requires the application of a correction method for restoring the accuracy. It seems that the fundamental Newton iteration from Section 3.2 can actually be used as a correction method. For instance, for small perturbations of the matrix

$$D \;=\; \begin{pmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & \lambda_2 \end{pmatrix},$$

it can be shown that the fundamental iteration still converges. However, for more general block diagonal matrices with triangular blocks, the details are quite technical and yet to be worked out.

Yet another direction for future investigations concerns the quadratic convergence. As a refinement of Lemma 1, we might replace D by a block diagonal matrix with entries $\Lambda_1, ..., \Lambda_p$. Instead of taking $B_{i,j} = \frac{M_{i,j}}{\lambda_j - \lambda_i}$, we then have to solve equations of the form

$$B_{i,j}\,\Lambda_j - \Lambda_i\,B_{i,j} \;=\; M_{i,j}.$$

If the Λ_i are sufficiently close to $\lambda_i\,\mathrm{Id}$, it might then be possible to adapt the fundamental iteration accordingly so as to achieve quadratic convergence for the strongly off diagonal part.

Bibliography

[1] G. Alefeld and J. Herzberger. *Introduction to interval analysis*. Academic Press, New York, 1983.

[2] L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and real computation*. Springer-Verlag, 1998.

[3] J. J. Dongarra, C. B. Moler, and J. H. Wilkinson. Improving the accuracy of computed eigenvalues and eigenvectors. *SIAM Journal on Numerical Analysis*, 20(1):23–45, 1983.

[4] G. H. Golub and F. Van Loan. *Matrix Computations*. JHU Press, 1996.

[5] S. Graillat and Ph. Trébuchet. A new algorithm for computing certified numerical approximations of the roots of a zero-dimensional system. In *Proc. ISSAC '09*, pages 167–174. ACM Press, 2009.

[6] J. van der Hoeven. Ball arithmetic. Technical report, HAL, 2009. http://hal.archives-ouvertes.fr/hal-00432152.

[7] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter. *Applied interval analysis*. Springer, London, 2001.

[8] F. Johansson. Arb: A c library for ball arithmetic. *ACM Commun. Comput. Algebra*, 47(3/4):166–169, 2014.

[9] U. W. Kulisch. *Computer Arithmetic and Validity. Theory, Implementation, and Applications*. Number 33 in Studies in Mathematics. de Gruyter, 2008.

[10] S. Miyajima. Fast enclosure for all eigenvalues in generalized eigenvalue problems. *Journal of Computational and Applied Mathematics*, 233(11):2994–3004, 2010.

[11] R. E. Moore. *Interval Analysis*. Prentice Hall, Englewood Cliffs, N.J., 1966.

[12] A. Neumaier. *Interval methods for systems of equations*. Cambridge university press, Cambridge, 1990.

[13] S. Rump and S. Graillat. Verified error bounds for multiple roots of systems of nonlinear equations. *Num. Algs.*, 54:359–377, 2010.

[14] S. M. Rump. Guaranteed inclusions for the complex generalized eigenproblem. *Computing*, 42(2):225–238, 1989.

[15] S. M. Rump. INTLAB - INTerval LABoratory. In Tibor Csendes, editor, *Developments in Reliable Computing*, pages 77–104. Kluwer Academic Publishers, Dordrecht, 1999. http://www.ti3.tu-harburg.de/rump/.

[16] S. M. Rump. Computational error bounds for multiple or nearly multiple eigenvalues. *Linear algebra and its applications*, 324(1-3):209–226, 2001.

[17] T. Yamamoto. Error bounds for computed eigenvalues and eigenvectors. *Numerische Mathematik*, 34(2):189–199, 1980.